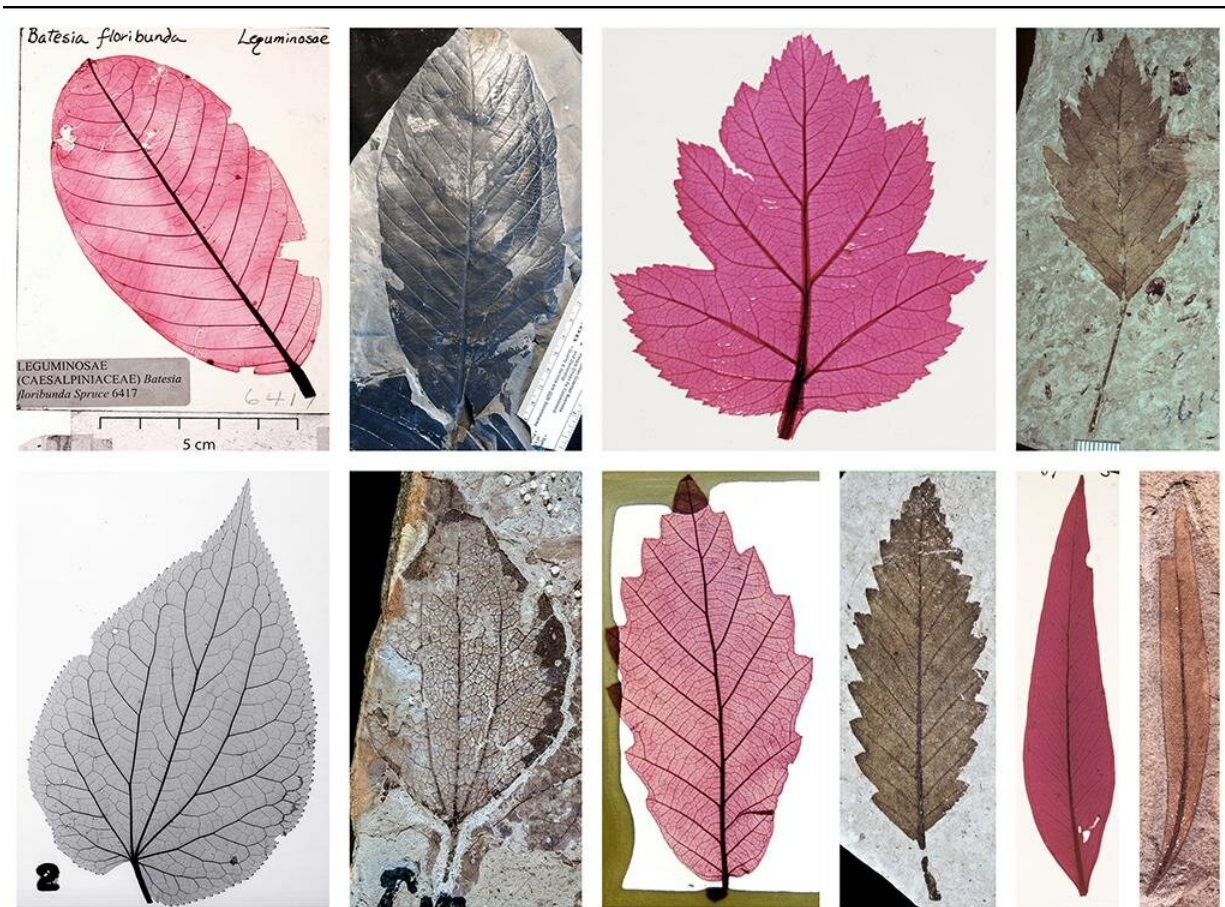# From museum to laptop: A visual leaf library is a new tool for identifying plants

March 15 2022

Selected pairs of modern and fossil leaves from a new, large dataset created by a Penn State-led team of scientists. Credit: Penn State

Fossil plants reveal the evolution of green life on Earth, but the most abundant samples that are found—fossil leaves—are also the most challenging to identify. A large, open-access visual leaf library developed by a Penn State-led team provides a new resource to help scientists recognize and classify these leaves.

"The complexity of leaves is off the charts, and the terminology we have to describe them is only the tiniest beginning of what is needed," said Peter Wilf, professor of geosciences at Penn State. "Researchers need much more accessible visual references to study what the differences are among the many plant groups, so we can put more of that into words. There are a lot of plant families that look superficially similar, and this collection provides an opportunity to see new patterns."

Studying fossil and modern leaves traditionally requires research visits to museum collections, which requires funding, planning and time for travel to several locations. More museums are putting leaf collections online, but often these images are low resolution, are hard to access in quantity, have uninformative filenames, or the leaves are photographed with other plant parts and labels that make rapid comparisons challenging, the scientists said.

The scientists combined images of modern and fossil leaves from several prominent collections, including several not previously online in any format, and spent thousands of hours formatting the data to create a single, merged, open-access dataset with standardized, easily searchable filenames and high-resolution images. They reported in *PhytoKeys* that the dataset is available from the Figshare Plus repository.

The dataset contains 30,252 images, including 26,176 images of cleared and x-rayed leaves and 4,076 fossil leaves. Cleared leaves are specimens that have been chemically bleached, stained and mounted on slides to reveal vein patterns. Each image represents a vouchered museum

specimen.

"What we have done here is to make this massive educational resource available to everyone by vetting and standardizing all these images from different legacy sources," Wilf said. "It took 15 years for us all to do that and convert all the filenames, but now you can have the whole package on your desktop with a single browser click. Every filename has the key information embedded, in the same order for rapid alpha-sorting: family, genus, species, and specimen number. The filenames can be rapidly searched in seconds for the item you are interested in and the images viewed using standard tools, such as the Windows search bar. All images are original resolution; nothing is downsampled."

The dataset is a potential resource not just to train students but also machine learning programs. Feeding vetted [training data] to learning algorithms allows them to better identify leaves and find important visual patterns that humans may have overlooked or been unable to see.

"For scientists studying botanical subjects, particularly fields such as paleobotany, these tools can most reliably be used to facilitate and multiply the impact of human expertise," said Jacob Rose, a doctoral student at Brown University, who worked closely with Wilf to create the dataset. His adviser, Thomas Serre, professor in computer science at Brown, also contributed. "Using these models as a starting point for an expert to either accept, reject or scrutinize further could soon prove to be a profound example of using technology to expand the value that is possible for a single scientist to produce as well as what is possible for us as a society to learn about the natural world, both in scale and precision."

Machine learning may be especially important for paleobotanists, who most often find isolated fossil leaves without seeds, fruit or flowers that could help identify plants. Further compounding the challenge, many of the individual fossils represent plants that are extinct.

The new dataset is a promising option for training machine learning because it contains examples of modern and fossil leaves vetted at least to the family level, a higher taxonomic classification that is the standard first target for fossil-leaf identification. The Fagaceae family, for example, includes beeches, chestnuts and oaks.

The dataset includes images from the Jack A. Wolfe and Leo J. Hickey contributions to the National Cleared Leaf Collection and the Scott Wing X-Ray collection at the Smithsonian National Museum of National History, Washington, D.C., and the Daniel I. Axelrod Cleared Leaf Collection at the University of California Museum of Paleontology, Berkeley. Also included are fossil images from various sites in North and South America. The largest contribution is from the Florissant Fossil Beds National Monument in Colorado.

"This database makes the information in these collections available to people around the world in a form that is easier to search than the original and more amenable to digital analyses," said Scott Wing, research geologist and curator of paleobotany at the Smithsonian. "We think the database will encourage new research and also open the museum collections to people."

  More information: Peter Wilf et al, An image dataset of cleared, x-rayed, and fossil leaves vetted to plant family for human and machine learning, *PhytoKeys* (2021). DOI: 10.3897/PhytoKeys.187.72350

Provided by Pennsylvania State University